ProfiT-HPC: Profiling Toolkit for High Performance Computing

Christian Boehme* und Igor Merkulow**

*Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen,
**Computational Health Informatics, Universität Hannover

8. Gauß-Allianz Statustagung (Erlangen)
08.-09.10.2018

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG)
Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)
KO 3394/14-1, OL 241/3-1, RE 1389/9-1, VO 1262/1-1, YA 191/10-1

Überblick

Projektpartner

Projektüberblick

Motivation

Sensibilisierung der Benutzer bezüglich der Performance ihrer Programme und bessere Ausnutzung der Ressourcen durch *Schulung*.

Ziele

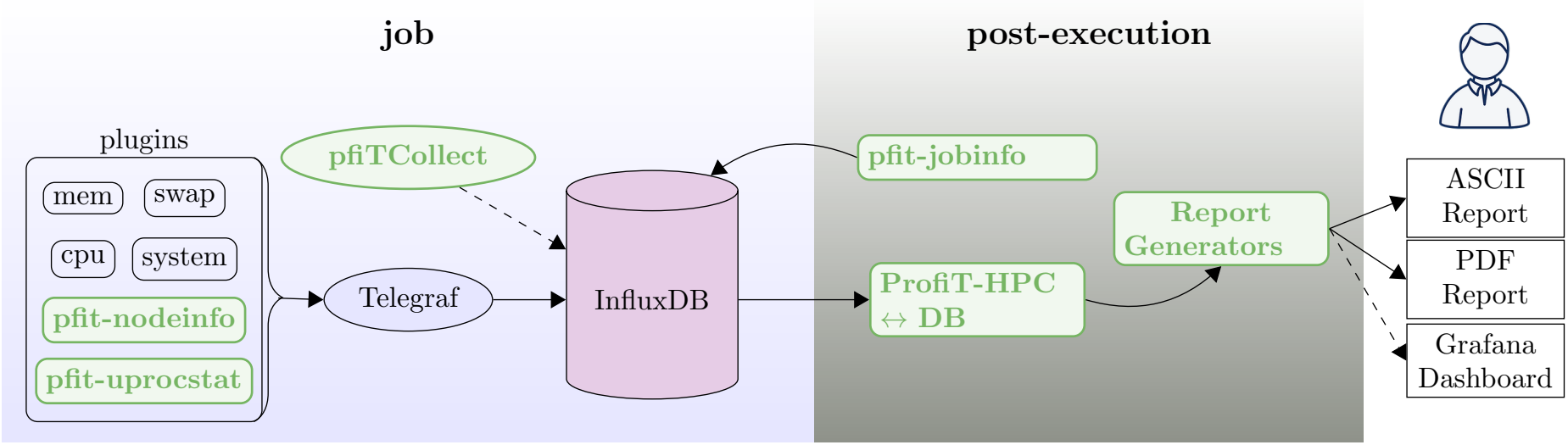
Automatisierte, Job-basierte Berichterstellung mit

- ausgewählten Metriken und Performanceindikatoren,
- potentiellen Problemen,
- Hinweisen zur Behebung dieser Probleme.

Vorarbeiten

goo.gl/wY6ybr	Results of a Survey concerning the Tier-2 and Tier-3 HPC-Infrastructure in Germany
goo.gl/LQ3mLt	Pre-Selection of the Metric Collection Tools
goo.gl/WvTXJc	Evaluation of Performance Monitoring Frameworks
goo.gl/VZbn9q	Concise Overview of Metrics and Tools

Architektur



Metriken

Knoten

- CPU
- Memory
- System

Prozess

- CPU times
- Memory
- IO

Filesystem

- Lustre
- BeeGFS
- Local

Andere

- Job
- GPU
- Netzwerk (Infiniband)

Telegraf - Plugins

Input-Plugins (übernommen und konfiguriert)

- cpu (CPU-Metriken),
- mem (Memory-Metriken),
- swap (Swap-Metriken),
- system (Metriken zur Auslastung des System).
- diskio (Metriken bzgl. Read/Write auf der lokalen Festplatte),
- lustre (Metriken bzgl. Read/Write)
- procstat (Prozessorientierte Metriken).

Telegraf - Plugins

Plugin-Eigenentwicklungen für zusätzliche Metriken

- Input-Plugins
  - GPU-Plugin
    - Metriken bzgl. vorhandener GPUs, z.B. benutzter GPU-Speicher,
    - basierend auf nvidia-smi (NVIDIA System Management Interface).
  - BeeGFS-Plugin
    - Metriken bzgl. Read/Write (Ops und Data),
    - basierend auf beegfs-ctl.
  - Infiniband-Plugin
    - Metriken bzgl. Daten und Anzahl von Paketen, welche über Infiniband-Ports verschickt oder empfangen wurden,
    - basierend auf perfquery.

Telegraf - Plugins

Plugin-Eigenentwicklungen für zusätzliche Funktionalität

- pfit-jobinfo (fügt JobId-Tag zu den Metriken hinzu),
- pfit-uprocstat (Prozess- und nutzerbasierte Metriken zur Unterstützung von shared nodes),
- pfit-nodeinfo (ermittelt Knoteninformationen).

PfiTCollect - Überblick

- Metrik-Kollektor (aktuell: Daten von ca. 50 Metriken),
- Geschrieben in C, erweiterbar bzgl. zusätzlicher Metriken,
- Keine Pluginstruktur, Metriken werden zentral im Quelltext implementiert,
- Aktuell keine Unterstützung für shared-nodes,
- Daten werden kompatibel zu Telegraf in InfluxDB gespeichert.

Telegraf oder PfiTCollect

Telegraf

+ Große Verbreitung
+ Unabhängig erweiterbar durch Plugins
- Schwergewichtig (27 MB Executable)
- Aus Quelltext schwer installierbar

PfiTCollect

+ Sehr leichtgewichtig (83 KB Executable)
+ Einfache Installation aus Quelltext
- Keine externe Entwickler-Community

Nutzer-Reports

- Text-Report
  - Abgelegt im Job-Verzeichnis
- PDF-Report
  - Abgelegt im Job-Verzeichnis
- Grafana-Dashboard
  - Per Link aus statischen Reports erreichbar

Text-Report - Überblick

Struktur des Berichts:

- General Job Information,
- Requested Resources,
- Node Description,
- Per Node Usage,
- Possible Problems and Recommendations.

Text-Report - Possible Problems

Hinweise zu Besonderheiten der Ressourcennutzung

- Große Unterschiede von angeforderten und genutzten Ressourcen
  - z.B. Nutzung von nur 10% der angeforderten Zeit
- Eventuell fehlerhafte Ressourcenanforderung
  - z.B. Anforderung von 8 Knoten mit jeweils 1 Kern, aber Rechnung mit 8 Threads auf nur einem Knoten
- auffällig hohe oder niedrige Belastung von Ressourcen
  - Nutzung von Swap,
  - hohe IO- oder Netzwerklast,
  - hohe Idle-Zeiten, ...

PDF-Report Überblick

- Job Überblick auf der ersten Seite
  - Informationen des Text-Reports mit visueller Unterstützung,
  - Unterschiedliche Diagrammarten.
- Zeitreihendarstellungen
- Funktionsprofile

Grafana Dashboards Überblick

- Interaktive und detaillierter Darstellung einzelner Jobs auf verschiedenen Ebenenen
  0. Job-Überblick
  1. Knotennutzung
  2. Prozessbasierte Statistiken
  3. Zeitreihen
  4. Ausführliche Informationen

Demo und Projekt-Paketierung

- Projekt-Demo wird als Docker-Lösung bereitgestellt
- Die Container sind modular einsetzbar
- InfluxDB-Container
  - Einschließlich Testdatensatz
- Grafana-Container
  - Einschließlich vorkonfigurierter Dashboards
- Selbstkonfigurierend
- Kann mit eigenen Datenquellen (Telegraf oder PfiTCollect) in Produktion überführt werden

DEMO

Offene Punkte

- Schnittstelle zwischen DB und Generatoren nicht abgeschlossen
- Entwicklung der Entscheidungskriterien für mehr Performance-Hinweise
  - Korrelation mehrerer Metriken
  - Erstellung der zugehörigen Best Practices
- Integration von automatisierten Funktionsprofilen
- Intensivierung der Tests des Prototyps
  - Möglichst auch bei Nicht-Partnern
  - Unterstützung durch Workshop
- Nutzung von Machine Learning

Deep Learning Integration

Aus ProfiT-HPC vergebene Masterarbeit

Autor: Azat Khuziyakhmetov

Thema: Anomaly detection of GPU utilization with neural networks

Zusammenfassung

- Nutzung einer globalen, Zeitreihenbasierten-DB hat sich bewährt
  - Skalierbare, modulare Lösung
  - Erweiterungen aus der Community integrierbar
- Pipeline zur Generation der verschiedenen Nutzeransichten funktionstüchtig
  - Verfügbar an allen beteiligten Zentren
  - Download in Vorbereitung
- Umfassende, automatisierte Ressourcen-Profile von HPC-Jobs